KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Robust forecasting of non-stationary time series

C. Croux, R. Fried, I. Gijbels and K. Mahieu

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1018

Robust forecasting of non-stationary time series

Croux C. *, Fried R. †, Gijbels I. ‡ and Mahieu K. §

September 6, 2010

Abstract

This paper proposes a robust forecasting method for non-stationary time series. The time series is modelled using non-parametric heteroscedastic regression, and fitted by a localized MM-estimator, combining high robustness and large efficiency. The proposed method is shown to produce reliable forecasts in the presence of outliers, non-linearity, and heteroscedasticity. In the absence of outliers, the forecasts are only slightly less precise than those based on a localized Least Squares estimator. An additional advantage of the MM-estimator is that it provides a robust estimate of the local variability of the time series.

Keywords: Heteroscedasticity, Non-parametric regression, Prediction, Outliers, Robustness

1 Introduction

This paper presents a *flexible* and *robust* forecasting technique for *non-stationary* and possibly *heteroscedastic* time series. Real data sets impose a simultaneous need for flexibility, robustness against outliers and the ability of coping with heteroscedasticity. Firstly, flexible modelling is adequate for time series with an underlying trend which does not lie in a predetermined family of parametric functions. Secondly, robustness should be involved to prevent outliers in the data from having a strong adverse effect on the predictions. Finally, the ability to handle heteroscedasticity

*Christophe Croux, ORSTAT, K.U.Leuven, Naamsestraat 69, B-3000 Leuven, Belgium; Tilburg University, Faculty Economics and Business Administration, P.O. Box 90153, 5000 LE Tilburg; Christophe.Croux@econ.kuleuven.be.

†Roland Fried, Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Deutschland; fried@statistik.tu-dortmund.de

‡Irène Gijbels, Mathematics Department, K.U.Leuven, Celestijnenlaan 200b - bus 2400, B-3001 Heverlee, Belgium; Irene.Gijbels@wis.kuleuven.be

§Koen Mahieu, ORSTAT, K.U.Leuven, Naamsestraat 69, B-3000 Leuven, Belgium; Koen.Mahieu@econ.kuleuven.be.

is of major importance, since restricting to homoscedasticity is too stringent for many real data sets. The forecasting procedure proposed here combines these three advantageous characteristics.

Most forecasting techniques found in the literature lack at least one of the aforementioned properties. Local polynomial regression (see e.g. Fan and Gijbels, 1996), for instance, is a flexible technique but not robust; furthermore, like local polynomial M-regression (Grillenzoni, 2009) and the weighted repeated median (Fried, Einbeck, and Gather, 2007), it does not explicitly take a possible change of the variance over time into account.

A key problem of many robust regression techniques is the estimation of the scale of the error term. Grillenzoni (2009) considered this scale to be estimated beforehand, whereas we estimate the scale and the trend simultaneously, using an S-estimator. S-estimators are known to have a high breakdown point but low efficiency. The efficiency of the procedure is then improved considerably by applying an MM-type estimator, initialized by the S-estimator. Additionally, since we estimate the error scale locally, pointwise prediction bounds can be constructed around the estimated trend, allowing the scale to change over time.

The rest of this paper is organized as follows. The underlying model and the procedure for the estimation of the trend are explained in Section 2. Section 3 describes an iterative procedure for computing the estimates. The choice of the kernel function, the bandwidth and the degree of the local polynomial approximating the trend are discussed in Section 4. In Section 5 a comparative simulation study is carried out. The proposed technique is illustrated by a real data example in Section 6. Some conclusions are stated in Section 7.

2 Model and estimation procedure

Consider the model

$$Y_t = m(t) + \sigma(t)\varepsilon_t$$

for a given time series Y_t, $t = 1, \ldots, T$. The signal m, capturing the underlying trend, and the scale σ are unknown functions of time. For simplicity, assume that the error terms ε_t are independent and distributed as $N(0, 1)$. To predict the value of Y_{T+1} we estimate the corresponding signal value $m(T+1)$. Additionally, we want to estimate the error scale $\sigma(T+1)$ at the same time point. In general, one could estimate the signal m and the scale σ in any other time point t_0, but we

focus on forecasting and choose $t_0 = T + 1$.

Let us approximate the signal m by a polynomial of degree p in a neighbourhood of t_0:

$$m(t) \approx m(t_0) + \beta_1(t - t_0) + \beta_2(t - t_0)^2 + \ldots + \beta_p(t - t_0)^p.$$

Consider the polynomial

$$\sum_{j=0}^{p} \beta_j (t - t_0)^j = \boldsymbol{\beta}' \boldsymbol{x}_{t,t_0},$$

with $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ and $\boldsymbol{x}_{t,t_0} = (1, (t - t_0), (t - t_0)^2, \ldots, (t - t_0)^p)'$, for each $t = 1, \ldots, T$. The signal $m(t_0)$ at time point t_0 is then estimated by

$$\hat{m}(t_0) = \hat{\beta}_0 = \hat{\boldsymbol{\beta}}' \boldsymbol{x}_{t_0,t_0}.$$

Typically, the parameter $\boldsymbol{\beta}$ is estimated by Local Polynomial Regression (LPR), minimizing a weighted Least Squares criterion, with weights decreasing as a function of the distance between t_0 and t. The LPR method is, however, not robust to outliers. Therefore we propose to estimate $\boldsymbol{\beta}$ by a local version of an MM-estimator (Yohai, 1987).

As a starting point for computing the MM-estimator we need an initial S-estimator for $\boldsymbol{\beta}$ and σ at time point t_0. The S-estimator $\hat{\boldsymbol{\beta}}_S$ for $\boldsymbol{\beta}$ is obtained by minimizing a scale function

$$\hat{\boldsymbol{\beta}}_S = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, S(\boldsymbol{\beta}), \tag{1}$$

where the scale function $S(\boldsymbol{\beta})$ is a localized M-scale defined as the solution of

$$\frac{1}{\sum_{t=1}^{T} K\left(\frac{t-t_0}{h}\right)} \sum_{t=1}^{T} \rho_0 \left(\frac{Y_t - \boldsymbol{\beta}' \boldsymbol{x}_{t,t_0}}{S(\boldsymbol{\beta})} \right) K\left(\frac{t - t_0}{h}\right) = b_0 \ . \tag{2}$$

Here, $\rho_0(\cdot)$ is a differentiable loss function controlling the influence of outliers, $K(\cdot)$ is a kernel function for downweighting observations that are far away from the time of interest t_0, and $h > 0$ is a bandwidth. The constant b_0 on the right-hand side of (2) is chosen to guarantee consistency at the normal distribution, i.e. $b_0 = \mathbb{E}(\rho_0(Z))$ with $Z \sim N(0, 1)$. Using a uniform kernel will give an equal positive weight to all observations in a time window around t_0, where the width of the window depends on the bandwidth h. The estimator β reduces then to a regular MM-estimator applied to the observations in the window, and has a breakdown point of $\varepsilon^* = \min(b_0/\rho_0(\infty), 1 - b_0/\rho_0(\infty))$ (Maronna *et al.*, 2006). The breakdown point measures the fraction of contamination in the time

Table 1: Asymptotic efficiencies at the normal distribution, of the MM-estimator for several values of the tuning constant c_1.

efficiency	80%	85%	90%	95%
c_1	3.14	3.44	3.88	4.68

window needed to change the local estimate by any given amount. Finally, the scale estimate at t_0 is defined as

$$\hat{\sigma} = S(\hat{\boldsymbol{\beta}}_S). \tag{3}$$

Once the initial local S-estimates $\hat{\boldsymbol{\beta}}_S$ and $\hat{\sigma}$ are obtained, we can improve the efficiency of $\hat{\boldsymbol{\beta}}_S$, while maintaining its breakdown point. This is done by applying an MM-step. For this purpose, denote by $\rho_1(\cdot)$ a second loss function satisfying $\rho_1 \leq \rho_0$. The local MM-estimate $\hat{\boldsymbol{\beta}}$ minimizes

$$\sum_{t=1}^{T} \rho_1 \left(\frac{Y_t - \boldsymbol{\beta}' \boldsymbol{x}_{t,t_0}}{\hat{\sigma}} \right) K \left(\frac{t - t_0}{h} \right), \tag{4}$$

with $\hat{\sigma}$ given by the initial S-estimator (3).

The loss functions ρ_0 and ρ_1 need to fulfill some conditions: they should be even, symmetric around 0, differentiable and nondecreasing in $|x|$ with $\rho_0(0) = \rho_1(0) = 0$. In order to guarantee robustness with a positive breakdown point, they need to be bounded. The biweight ρ-function is frequently used in S- and MM-estimation and is defined as

$$\rho_c(x) = \begin{cases} \left(1 - \left(1 - (x/c)^2\right)^3\right) & \text{if } |x| \leq c \\ 1 & \text{otherwise.} \end{cases} \tag{5}$$

The tuning constant c determines the robustness. We use the biweight ρ-function for both the S- and the MM-estimator, taking $\rho_0 = \rho_{c_0}$ and $\rho_1 = \rho_{c_1}$. To achieve high robustness of the S-estimator we choose $c_0 = 1.5476$, resulting in a 50% breakdown point in case of an uniform kernel K. The larger c_1, the more efficient is the MM-estimator at the normal. Common values of c_1 corresponding to certain asymptotic efficiencies at the normal distribution are given in Table 1. We take $c_1 = 3.88$, corresponding to an efficiency of 90%. The selection of the kernel function K, of the bandwidth h, and of the polynomial degree p will be discussed in Section 4.

3 Computation

This section describes the practical computation of the proposed estimator. The procedure can be divided into three main steps. Firstly, an initial estimate of $\boldsymbol{\beta}$ and σ is needed to start the computation of the S-estimator. Secondly, the S-estimate of $\boldsymbol{\beta}$ and σ is computed. Finally, the efficiency of the estimation of $\boldsymbol{\beta}$ is improved by means of an MM-estimate, while the S-estimate of σ is kept constant.

Step 1: An initial estimate of $\boldsymbol{\beta}$ is obtained by local linear least absolute deviation (LAD) regression. It is particulary useful in this context, since it does not require an auxiliary scale estimate. It is defined as

$$\hat{\boldsymbol{\beta}}_0 = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{t=1}^{T} |Y_t - \boldsymbol{\beta}' \boldsymbol{x}_{t,t_0}| K^h_{t,t_0}, \tag{6}$$

where $K^h_{t,t_0} = K\left(\frac{t-t_0}{h}\right)$, to shorten the notation. Computing (6) is easy, since it is equivalent to an unweighted LAD regression on the data $\left(x_{t,t_0}K^h_{t,t_0}, Y_t K^h_{t,t_0}\right)$, for $t = 1, \ldots, T$.

As an initial estimate of σ we suggest the "weighted median absolute deviation from zero" of the LAD regression residuals. More precisely, we compute the weighted median of the absolute values of the residuals $|Y_t - \boldsymbol{\beta}' \boldsymbol{x}_{t,t_0}|$, with associated weights K^h_{t,t_0}, for $t = 1, \ldots, T$. Computation of a weighted median is standard, e.g. Fried *et al.* (2007). Denote the resulting initial estimate of σ by $S_0(\hat{\boldsymbol{\beta}}_0)$.

Step 2: The second step uses an iterative computation scheme for the S-estimates of $\boldsymbol{\beta}$ and σ, starting from the initial estimates derived in the first step. Denote the derivative of $\rho_0(\cdot)$ by $\psi_0(\cdot)$. Taking the derivative of (2) with respect to the vector $\boldsymbol{\beta}$ leads to the following set of estimation equations:

$$\sum_{t=1}^{T} \psi_0\left(\frac{Y_t - \boldsymbol{\beta}' \boldsymbol{x}_{t,t_0}}{S(\boldsymbol{\beta})}\right) \left(\frac{-S(\boldsymbol{\beta})\boldsymbol{x}_{t,t_0} - (Y_t - \boldsymbol{\beta}' \boldsymbol{x}_{t,t_0}) \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}}{S(\boldsymbol{\beta})^2}\right) K^h_{t,t_0} = \mathbf{0}.$$

Isolating $\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ and setting it to zero yields the set of first order conditions:

$$\sum_{t=1}^{T} \psi_0\left(\frac{Y_t - \boldsymbol{\beta}' \boldsymbol{x}_{t,t_0}}{S(\boldsymbol{\beta})}\right) K^h_{t,t_0} \boldsymbol{x}_{t,t_0} = 0,$$

which can be rewritten as

$$\sum_{t=1}^{T} w_t(\boldsymbol{\beta}) K^h_{t,t_0} (Y_t - \boldsymbol{\beta}' \boldsymbol{x}_{t,t_0}) \boldsymbol{x}_{t,t_0} = 0, \tag{7}$$

where the weights $w_t(\boldsymbol{\beta})$ are defined as

$$w_t(\boldsymbol{\beta}) = \psi_0\left(\frac{Y_t - \boldsymbol{\beta}'\boldsymbol{x}_{t,t_0}}{S(\boldsymbol{\beta})}\right) / \left(\frac{Y_t - \boldsymbol{\beta}'\boldsymbol{x}_{t,t_0}}{S(\boldsymbol{\beta})}\right). \tag{8}$$

The weights w_t downweight observations with large residual values. Note that the weights are equal to one if ψ_0 is the identity function, corresponding to an ordinary least squares fit. If we assume that the residual values are known, equation (7) is simply the first order equation of a weighted least squares estimator, with weights equal to the product of the kernel weights K^h_{t,t_0} and the residual weights w_t. Hence, we can express the solution of (7) as

$$\hat{\boldsymbol{\beta}} = (X'W_{\boldsymbol{\beta}}X)^{-1}X'W_{\boldsymbol{\beta}}Y,$$

with $Y = (Y_1, \ldots, Y_T)'$ the vector of all observations, and where the design matrix X and the weight matrix $W_{\boldsymbol{\beta}}$ are given by

$$X = \begin{pmatrix} 1 & (1-t_0) & \cdots & (1-t_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (T-t_0) & \cdots & (T-t_0)^p \end{pmatrix}, \quad W_{\boldsymbol{\beta}} = \begin{pmatrix} w_1(\boldsymbol{\beta})K^h_{1,t_0} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_T(\boldsymbol{\beta})K^h_{T,t_0} \end{pmatrix}.$$

Given the initial values $\hat{\boldsymbol{\beta}}_0$ and $S_0(\hat{\boldsymbol{\beta}}_0)$ from step one, one computes iteratively

$$\hat{\boldsymbol{\beta}}_i = (X'W_{\hat{\boldsymbol{\beta}}_{i-1}}X)^{-1}X'W_{\hat{\boldsymbol{\beta}}_{i-1}}Y, \; i = 1, 2, \ldots. \tag{9}$$

Note that in the i-th iteration, a scale $S(\hat{\boldsymbol{\beta}}_{i-1})$ that fulfills (2) is needed for the computation of the residual weights (8). This condition can be rewritten as

$$\frac{S(\hat{\boldsymbol{\beta}}_{i-1})}{b_0 \sum_{t=1}^T K^h_{t,t_0}} \sum_{t=1}^T \rho_0\left(\frac{Y_t - \hat{\boldsymbol{\beta}}'_{i-1}\boldsymbol{x}_{t,t_0}}{S(\hat{\boldsymbol{\beta}}_{i-1})}\right) K^h_{t,t_0} = S(\hat{\boldsymbol{\beta}}_{i-1}),$$

and suggests the following iterative scheme for the scale:

$$S_k(\hat{\boldsymbol{\beta}}_{i-1}) = \frac{S_{k-1}(\hat{\boldsymbol{\beta}}_{i-1})}{b_0 \sum_{t=1}^T K^h_{t,t_0}} \sum_{t=1}^T \rho_0\left(\frac{Y_t - \hat{\boldsymbol{\beta}}'_{i-1}\boldsymbol{x}_{t,t_0}}{S_{k-1}(\hat{\boldsymbol{\beta}}_{i-1})}\right) K^h_{t,t_0}, \; k = 1, 2, \ldots \tag{10}$$

using an initial value $S_0(\hat{\boldsymbol{\beta}}_{i-1})$. For $i > 1$, we take the scale estimate obtained in iteration $i-1$ for $\hat{\boldsymbol{\beta}}$ as the initial value $S_0(\hat{\boldsymbol{\beta}}_i)$.

Schematically, the S-estimator is calculated as follows:

1. Calculate $\hat{\boldsymbol{\beta}}_0$ by local linear least absolute deviation regression and $S_0(\hat{\boldsymbol{\beta}}_0)$ by weighted median absolute deviation from zero.

2. Calculate $S(\hat{\boldsymbol{\beta}}_0)$ by iterating equation (10).

3. For $i = 1, 2, \ldots$:

   a) Calculate $\hat{\boldsymbol{\beta}}_i$ using equation (9) with initial values $\hat{\boldsymbol{\beta}}_{i-1}$ and $S(\hat{\boldsymbol{\beta}}_{i-1})$ for the scale.

   b) Calculate $S(\hat{\boldsymbol{\beta}}_i)$ by iterating equation (10) with initial value $S(\hat{\boldsymbol{\beta}}_{i-1})$ up to convergence.

After achieving convergence, say in step ℓ, we obtain the S-estimates $\hat{\boldsymbol{\beta}}_S = \hat{\boldsymbol{\beta}}_\ell$ and $\hat{\sigma} = S(\hat{\boldsymbol{\beta}}_\ell)$.

Step 3: The MM-estimate is computed as the minimizer of (4). The corresponding first order condition can be solved with an iterative scheme of the same form as (9), except for the scale, which is held fixed and equal to the scale $\hat{\sigma}$ obtained in the second step. More precisely, the final MM-estimator of the regression coefficients is computed by iterating

$$\hat{\boldsymbol{\beta}}_i = (X'\tilde{W}_{\hat{\boldsymbol{\beta}}_{i-1}}X)^{-1}X'\tilde{W}_{\hat{\boldsymbol{\beta}}_{i-1}}Y, \quad i = 1, 2, \ldots. \tag{11}$$

Here, the weight matrix $\tilde{W}_{\boldsymbol{\beta}}$ is defined by

$$\tilde{W}_{\boldsymbol{\beta}} = \begin{pmatrix} \tilde{w}_1(\boldsymbol{\beta})K^h_{1,t_0} & & 0 \\ & \ddots & \\ 0 & & \tilde{w}_T(\boldsymbol{\beta})K^h_{T,t_0} \end{pmatrix},$$

with $\tilde{w}_t(\boldsymbol{\beta}) = \psi_1\left(\frac{Y_t - \boldsymbol{\beta}'\boldsymbol{x}_{t,t_0}}{\hat{\sigma}}\right) / \left(\frac{Y_t - \boldsymbol{\beta}'\boldsymbol{x}_{t,t_0}}{\hat{\sigma}}\right)$ and ψ_1 the derivative of ρ_1. The starting value $\hat{\boldsymbol{\beta}}_0$ for the iterative scheme (11) is the S-estimator $\hat{\boldsymbol{\beta}}_S$ from step two. Extensive simulations indicated that the algorithm converges in practically all cases we considered.

4 Kernel function, bandwidth and polynomial degree

Generally, kernel smoothing techniques use symmetric, unimodal kernel functions. Moving the corresponding window one time point further into the future only has a small effect on the estimate $\hat{m}(t_0)$. Therefore, if a reasonable bandwidth is used, the sequence of fitted values gives a smooth curve. In a forecasting context, however, only observations left of the target point t_0, corresponding

to past values, are available. Moreover, it is intuitive to put the largest weights on the observations just before the time point at which we want to make a prediction. Hence, asymmetric kernel functions should be used. For asymmetric kernels, the observation which has entered the time window most recently gets the largest weight, so that the sequence of fitted values (the forecasts) will be less smooth compared to the ones obtained using a symmetric kernel. The precise shape of the kernel function K is of less importance. We use an asymmetric exponential kernel function $K(x) = \exp(x)I_{\{x<0\}}$ as in Gijbels *et al.* (1999), where local polynomial regression is linked to Holt-Winters forecasting.

As opposed to the shape of the kernel function, the selection of the bandwidth h in local polynomial regression is crucial and deserves particular attention. The trade-off between bias and variance of the estimator $\hat{m}(t)$ is well-known: the bias increases with h, while the variance decreases. In the context of forecasting, h determines the influence of past observations on the prediction. Small values of h yield forecasts that mainly rely on the most recent observations. Large values of h correspond to more slowly varying forecasts.

The optimal bandwidth can be chosen as the one minimizing the mean of the squared one-step ahead forecast errors (MSFE). Given a bandwidth h, the method is applied to forecast Y_t, using all previous observations. This results in a one-step ahead forecast error e_t, for each $t = t_{\min}, \ldots, T$. Here $t_{\min}$ is a small value, but at least larger than $p + 2$. In our application we take $t_{\min} = 21$, since we believe that at least 20 observations are needed to make a reliable forecast. To predict the target value Y_{T+1}, one could use the bandwidth that minimizes the mean of the squared forecast errors $e_{t_{\min}}, \ldots, e_T$. Note that bandwidth selection by minimizing MSFE is equivalent to bandwidth selection by cross-validation (Heng-Yan Leung, 2005), since the value of the time series at the time point of interest is not used in its prediction.

By using the MSFE, however, one implicitly assumes homoscedastic error terms that are free of outliers. Our method, in contrast, is designed to cope with heteroscedasticity and outliers. Hence, the approach for bandwidth selection should share these properties, and thus we need to adapt the criterion. First, to remove the effect of the changing variance, we standardize each one-step ahead forecast error by dividing it by a local scale estimate $\hat{\sigma}(t)$. This local scale estimate is obtained as a byproduct of the proposed forecasting method. When predicting Y_{t+1}, equation (3) provides an estimate for $\sigma(t)$. Second, to limit the effect of possible outliers, we apply a trimmed mean of

the squared standardized one-step ahead forecast errors as the criterion for bandwidth selection. The percentage of trimming α can be specified by the user, depending on the application. In this paper, we set $\alpha = 0.2$, trimming the largest 20% of the squared standardized forecast errors. To summarize, for predicting Y_{T+1} we select the bandwidth minimizing

$$\frac{1}{\lfloor (1-\alpha)\bar{T} \rfloor} \sum_{i=1}^{\lfloor (1-\alpha)\bar{T} \rfloor} \left(\frac{e_t}{\hat{\sigma}(t)} \right)^2_{(i)}, \tag{12}$$

with $\bar{T} = T - t_{\min} + 1$ and where $(e_t/\hat{\sigma}(t))^2_{(i)}$ is the i'th term in the ordered vector of squared standardized forecast errors for time point $t_{\min}$ to T.

The selection of the degree of the polynomial used in the signal approximation is less important than the selection of the bandwidth. Large values of the degree p yield a high variability, but have an advantageous effect on the bias. Moreover, the larger p, the higher the computational cost. We use polynomials of degree $p = 1$, i.e. the signal or trend is supposed to be locally linear. This choice is often suggested in the literature (see for instance Fan and Gijbels, 1996).

5 Simulation study

In this section we carry out a simulation study in order to compare our new forecasting approach to methods that already exist. The following four methods are included in this comparison: the regular Local Polynomial Regression (LPR), the Weighted Repeated Median (WRM) technique of Fried, Einbeck, and Gather (2007), the local polynomial M-smoother of Grillenzoni (2009), and finally the local polynomial MM-estimator proposed in this paper. In each method, we use the asymmetric exponential kernel, defined as $K(x) = \exp(x) I_{\{x<0\}}$. The M-procedure is implemented as in Grillenzoni (2009), using the Huber ψ-function with tuning constant $k = 1.345$, and as auxiliary scale estimator the weighted median absolute deviation from zero of the residuals.The computation of the WRM is based on the R-package *robfilter*, developed by Fried *et al.* (2010). All local polynomials are of degree $p = 1$, which is a fair choice with respect to the linear WRM.

The forecasting methods should be applicable to time series with non-linear trends $m(t)$ and time varying variance $\sigma^2(t)$. By adding patterns of outliers, the robustness of the procedures is investigated. We generate time series from the model

$$Y_t = m(t) + \sigma(t)\varepsilon_t, \tag{13}$$

for $t = 1, \ldots, T$. The signal $m(t)$ is a sinusoidal function defined as $m(t) = 12.5 \sin(t\pi/200)$, similar to Fried, Einbeck, and Gather (2007). We set $\sigma(t) = m(t)/6$ to bring in heteroscedasticity. We consider three settings for the noise term ε_t:

1. *clean*: the noise term ε_t is an i.i.d. standard normal process.

2. *cont5%*: the noise term ε_t is the sum of an i.i.d. standard normal process and an outlier generating process η_t. At every time point there is a 5% probability that the observation is shifted upwards by an amount of $8\sigma(t)$, creating an additive outlier. The size of the outlier is proportional to the scale of the series. Hence, $\mathbb{P}(\eta_t = 0) = 0.95$ and $\mathbb{P}(\eta_t = 8\sigma(t)) = 0.05$.

3. *cont10%*: same as *cont5%*, but with $\mathbb{P}(\eta_t = 0) = 0.9$ and $\mathbb{P}(\eta_t = 8\sigma(t)) = 0.1$ corresponding to about 10% outliers.

In each of the settings, we consider time series of length 100. To take different levels of linearity into account, we compute one-step ahead predictions from $t = 50$ to $t = 100$. Obviously, the signal function $m(t)$ is more linear at time point 50 than at time point 100.

A bandwidth needs to be selected for each prediction. For the proposed MM-estimator, this is done as described in Section 4, by minimizing the trimmed mean squared standardized one-step ahead forecast errors. For the other estimators, a similar bandwidth selection method is used. The only difference is the scale, which is used to standardize the forecast errors. For LPR we estimate $\sigma(t)$ by a weighted standard deviation of the residuals of the local polynomial fit. For the WRM and the local polynomial M-estimator we use a weighted median absolute deviation from zero, where the weights are equal to the kernel weights used for prediction. The optimal bandwidth is obtained by a grid search, where a suitable grid is found by preliminary experiments.

Forecast Performance

We generate 1000 time series from the model, for each of the three simulation settings. Note that for each method, in each setting and for every time point for which a forecast is made, a specific bandwidth is selected. In Figure 1 we plot one single series generated according to the simulation scheme *cont5%*. We also add the one-step ahead predictions made by the LPR method (squares) and the MM procedures (crosses). We see that the predictions made by the LPR method are upwards biased at the time points right after the occurrence of the outliers, while the MM

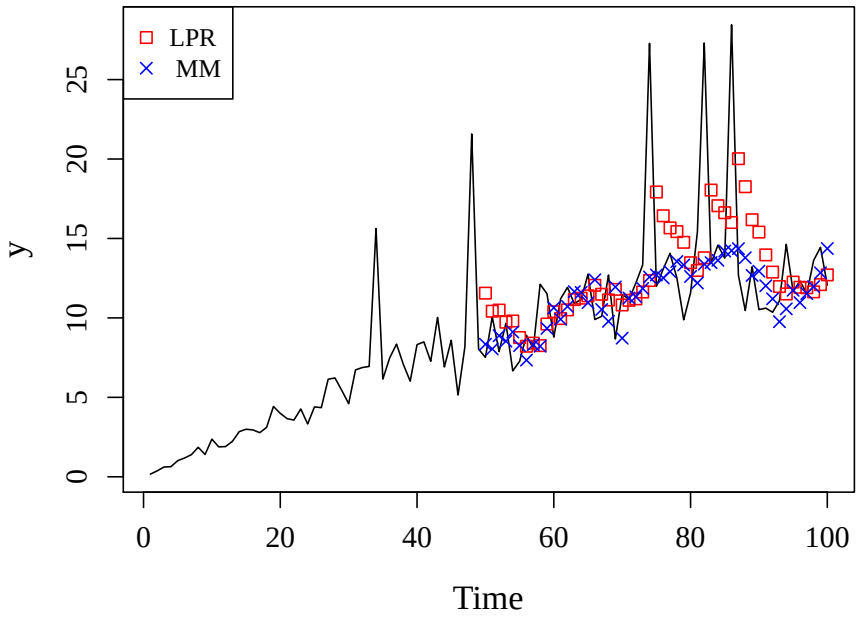

Figure 1: A simulated time series with 5% outliers, with one-step ahead predictions by LPR and MM, from time point 50 up to 100.

remains more or less unaffected. As expected, the outliers themselves are not well predicted, by neither of the methods.

By averaging over the 1000 simulation runs, we obtain a mean squared forecast error. Since we do not want that outliers affect the performance criterion, we report a 20% trimmed mean squared forecast error (TMSFE) instead. Alternatively, one could compute medians or choose other trimming percentages (larger than the true level of outliers), but this is not affecting the relative performance of the different procedures. Ignoring outliers in the evaluation of the forecast performance is reasonable, since outliers are difficult to predict and should be treated differently. Tables 2 and 3 present the TMSFE for prediction of Y_{50}, respectively Y_{100}, for the four different methods we consider.

From Table 2, it can be seen that in the clean data setting, LPR and MM perform almost equally well. The WRM is the least efficient in the absence of outliers. When the time series contains outliers, as in the two other simulation schemes, LPR loses its accuracy, whereas the MM is much less affected. The forecast performance of the MM, as measured by the TMSFE, is much better than for the two other robust procedures, WRM and M. For small amounts of contamination

Table 2: 20% right trimmed means of the squared one-step ahead forecast errors at time point 50, from 1000 simulation runs.

	LPR	WRM	M	MM
clean	1.044	1.211	1.158	1.087
cont5%	2.202	1.713	1.598	1.370
cont10%	3.960	2.377	3.036	1.978

Table 3: 20% right trimmed means of the squared one-step ahead forecast errors at time point 100, from 1000 simulation runs.

	LPR	WRM	M	MM
clean	2.486	2.827	2.410	2.560
cont5%	4.648	3.905	3.476	3.320
cont10%	6.482	5.198	4.633	3.619

(5%), the M-procedure still works relatively well, whereas for the larger percentages of outliers, it loses a lot of its robustness. The differences between the methods become more pronounced when the percentage of outliers increases from 5% to 10%.

In Table 3, the prediction is made at $t = 100$, where the signal is more non-linear. This makes the prediction exercise harder, as is witnessed by the larger values of the TMSFE in Table 3 compared to Table 2. The relative performance of the different methods is, however, still about the same. The MM-procedure is close to the LPR in the absence of outliers, and outperforms all other considered methods for the outlier generating simulation schemes.

6 Real data example

To demonstrate the practical relevance of the proposed forecasting method, we apply it to a real time series example. It concerns 150 daily maximal temperature values in Dresden, Germany, from 17 November 2006 to 15 April 2007. The data are plotted in Figure 2. At first sight, there

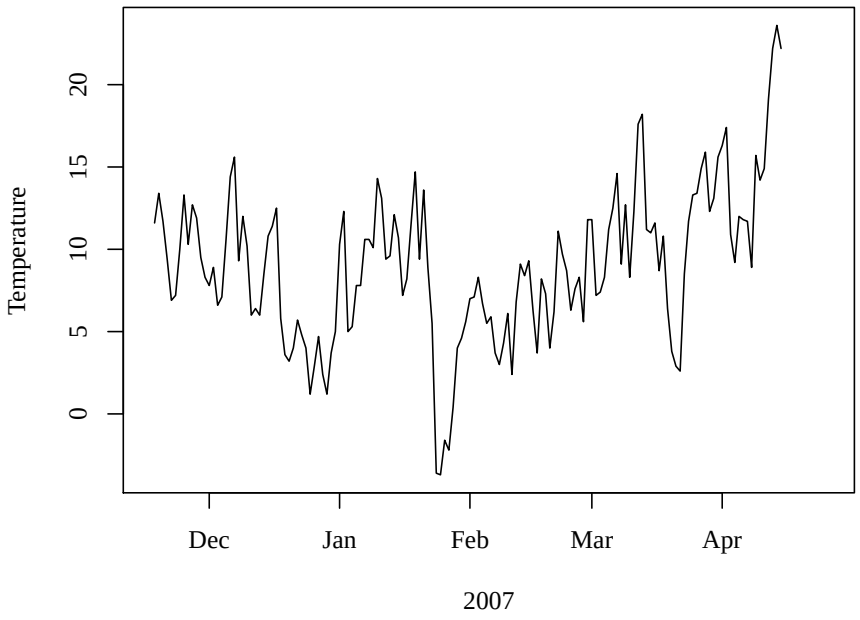

Figure 2: Temperature data in Dresden, Germany.

are only a few not very large outliers around 25 January and 22 March 2007, but there might be other smaller outliers in the series.

We proceed as in the simulation study. To predict the value y_t, the first $t-1$ observations are used. Predictions are made from 25 February 2007 up to 15 April 2007, i.e. the last 50 days. Each day a prediction is made, the bandwidth is selected according to the procedure described in Section 4. We take a closer look at what happens on 25 March 2007, since outliers occurred on the days before. The target functions for the bandwidth selection procedure for the prediction on 25 March are depicted in Figure 3. In the neighbourhood of the minimum, the curves are quite flat, especially for the MM-method, indicating a kind of robustness against bandwidth misspecification. The selected bandwidths for 25 March 2007 are listed in the following table.

	LPR	WRM	M	MM
h	12	17	12	22

The effect of the outliers appears through the smaller bandwidth that is selected for LPR and M.

To measure the forecast performance, we compute the 20% right trimmed means of the squared forecast errors over the whole prediction period. As such, the performance criterion we use is also robust with respect to outliers. Results are reported in Table 4, for the four different methods we

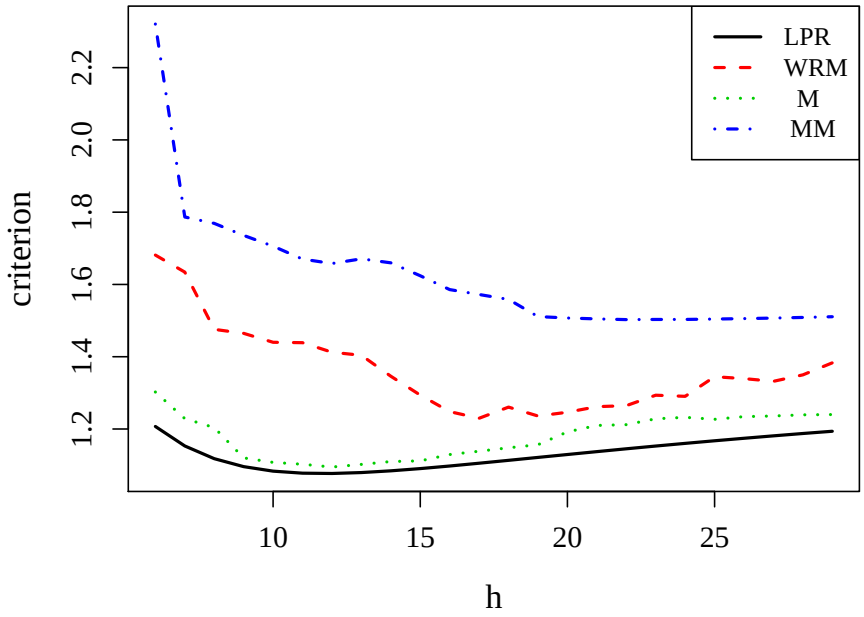

Figure 3: Target functions for the selection of the bandwidth on 25 March 2007.

consider. We see that the forecast performance of the different methods is quite similar, with the MM-procedure resulting in the smallest TMSFE for this particular data set.

Table 4: The trimmed mean squared forecast error (TMSFE) for the prediction of the daily temperature data from 25 February 2007 to 15 April 2007.

	LPR	WRM	M	MM
TMSFE	7.37	8.31	7.03	6.88

7 Conclusions

While the practical relevance of robust forecasting methods is without doubt, they received little attention in the robustness literature. Exceptions are Gelper *et al.* (2009) and Cipra (1992), who robustified the simple Holt-Winters forecast procedure and obtained a computationally fast ad-

hoc procedure. Robust smoothing of time series received more attention. Recently, an M-based smoothing approach was proposed by Grillenzoni (2009) and weighted repeated median smoothing by Fried *et al.* (2007). We investigated the performance of these two smoothing methods in a forecasting context, and included them in the simulation study as benchmark procedures.

In this paper we develop a new robust time series forecasting methodology for non-stationary time series. It allows for heteroscedasticity in the data and remains reliable in the presence of outliers. The technique is nonparametric and combines local polynomial regression and MM-estimation. It provides a local scale estimate as a byproduct. Our simulation study indicates that the new forecasting method outperforms other methods such as Local Polynomial regression, Weighted Repeated Median smoothing and local polynomial M-estimation in the presence of outliers and heteroscedasticity, while it still achieves comparable performance results in an uncontaminated setting.

Since the estimation procedure involves a local scale of the one-step ahead forecast errors, one could use this scale estimate for the construction of prediction intervals. A drawback of these scale estimates is that they suffer from a finite sample bias. Future research is needed to construct correctly sized robust prediction intervals. Furthermore, while computing the forecasting with the algorithm outlined in section 3 is fast, the selection of the optimal bandwidth is more time consuming. The selection of the bandwidth is currently redone at every time point, since the bandwidth may change over time. The alternative of updating the optimal bandwidth from the previous time point would be an interesting topic for further research.

References

Cipra, T. (1992). Robust exponential smoothing. *Journal of Forecasting*, **11**(1), 57–69.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall/CRC, 2003 edition.

Fried, R., Einbeck, J., and Gather, U. (2007). Weighted repeated median smoothing and filtering. *Journal of the American Statistical Association*, **102**(480), 1300–1308.

Fried, R., Schettlinger, K., and Borowski, M. (2010). *robfilter: Robust Time Series Filters*. R package version 2.6.1.

Gelper, S., Fried, R., and Croux, C. (2009). Robust forecasting with exponential and Holt-Winters smoothing. *Journal of Forecasting*, **29**, 285–300.

Gijbels, I., Pope, A., and Wand, M. P. (1999). Understanding exponential smoothing via kernel regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **61**(1), 39–50.

Grillenzoni, C. (2009). Robust non-parametric smoothing of non-stationary time series. *Journal of Statistical Computation and Simulation*, **79**(4), 379–393.

Heng-Yan Leung, D. (2005). Cross-validation in nonparametric regression with outliers. *The Annals of Statistics*, **33**(5), 2291–2310.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust statistics: Theory and Methods*. Wiley.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, **15**(2), 642–656.